心理学报 2023, Vol. 55, No. 1, 142−158
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2023.00142

认知诊断模型 Q 矩阵修正：完整信息矩阵的作用*

刘彦楼 [1] 吴琼琼 [2]

([1] 曲阜师范大学教育大数据研究院; [2] 曲阜师范大学心理学院, 山东 济宁 273165)

摘 要 Q 矩阵是 CDM 的核心元素之一, 反映了测验的内部结构和内容设计, 通常由领域专家根据经验进行主观界定, 因此需要对可能存在的错误进行修正。本研究提出了一种新的 Q 矩阵修正方法——基于完整经验交叉相乘信息矩阵的 Wald-XPD 方法。采用 Monte Carlo 模拟检验了新方法的表现, 并与同类方法进行了比较。研究表明：新开发的 Wald-XPD 方法在 Q 矩阵恢复率、保留正确标定属性的比例以及修正错误标定属性的比例这 3 个主要指标上均有较好的表现, 且整体上优于其他方法, 尤其是在修正错误标定的属性方面。通过实证数据展示了 Wald-XPD 方法在 Q 矩阵修正中的良好表现。总之, 本研究为 Q 矩阵修正提供了有效的方法。

关键词 认知诊断模型, Q 矩阵, XPD 矩阵, Wald 检验
分类号 B841

1 引言

经典心理测量理论及项目反应理论采用单一的测验分数来描述被试在某个阶段的学习效果。作为新一代心理测量理论, 认知诊断(cognitive diagnosis)的主要目的是提供关于被试的多维、细粒度潜在特质(如知识、认识过程、技能、策略、人格特质或心理障碍等, 统称为属性)的诊断性评价信息, 认知诊断模型(cognitive diagnostic model, CDM)是研究者为了实现以上主要目的而提出的一类离散潜变量模型(Rupp et al., 2010)。目前, CDM 已广泛应用于心理、教育、精神病理学等领域(Sorrel et al., 2016)。

Q 矩阵是 CDM 的核心元素之一, 定义了测验所测属性与项目之间的对应关系(Tatsuoka, 1990), 它不仅决定着测验的内部结构, 也关系到认知诊断结果的准确性。正确设定的 Q 矩阵是获得准确的模型参数估计和被试分类的关键因素(Nájera et al., 2020), 错误设定的 Q 矩阵会产生很多不良的影响, 如降低模型参数估计准确性、导致较差的模型−数据拟合、导致错误的属性估计和被试分类等(Chiu, 2013; de la Torre, 2009; Rupp & Templin, 2008)。CDM 中获取 Q 矩阵的方法主要是由领域专家根据经验构建(Sorrel et al., 2016), 但这种方法包含一定的主观性。实践中, 原始 Q 矩阵有较大可能包含一些错误设定(Rupp & Templin, 2008), 如何修正原始 Q 矩阵中可能存在的错误是研究者面临的重要理论与现实问题。

为了获得正确设定的 Q 矩阵, 国内外研究者提出了多种修正方法(李佳 等, 2021)。根据是否采用参数化的 CDM 描述 Q 矩阵与观察作答数据之间的关系, 可以将 Q 矩阵修正方法分为两类：参数化和非参数化的修正方法, 前者需要参数化 CDM 的参与, 后者不需要。例如, 欧氏距离法(Chiu, 2013)、海明距离(汪大勋, 高旭亮, 韩雨婷 等, 2018)、交差方法(intersection and difference; Wang et al., 2018)

收稿日期: 2022-03-09
* 国家自然科学基金青年项目(31900794)、山东省自然科学基金项目(ZR2019BC084)资助。
吴琼琼为共同第一作者。
通信作者: 刘彦楼, E-mail: liuyanlou@163.com

等属于非参数化的修正方法。一般而言，非参数化方法比较的是理想反应与观察作答反应之间的拟合，从而达到修正 **Q** 矩阵的目的。在非参数化方法中，理想反应大多都是在限制条件较为严格的情景下获得的，例如，限定所有项目只适用于某个或某几个特殊的(亦称，简化的)CDM。换言之，非参数化的 **Q** 矩阵修正方法具有样本量要求小、易实现等优点，但严格的前提条件限制了这些方法的拓展性及实用性。参数化 **Q** 矩阵修正方法是在参数化模型框架下，使用各种统计量估计出最能拟合观察数据的 **Q** 矩阵。在特殊的 CDM 框架下，如 DINA、DINO、R-RUM 等(de la Torre, 2011)，研究者开发的参数化修正方法主要有：δ 法(de la Torre, 2008)、γ 法(涂冬波 等, 2012)、S 统计量方法(Liu et al., 2012)、迭代修正序列搜索(iterative modified sequential search; Terzi & de la Torre, 2018)、RMSEA 统计量(Kang et al., 2019)、加权残差 R 法(Yu & Cheng, 2020)、最优反应分布纯度方法(李佳 等, 2022)等。在饱和 CDM 框架下(如，G-DINA, generalized deterministic input noisy output “and” gate; de la Torre, 2011)的参数化 **Q** 矩阵修正方法主要包括：GDI (G-DINA discrimination index)方法(de la Torre & Chiu, 2016)、残差方法(Chen, 2017)、iJSD (iterative Jensen-Shannon divergence)方法以及 iGDI (iterative GDI)方法(Terzi, 2017)、TLP (truncated L_1 penalty function)方法(Xu & Shang, 2018)、相对拟合统计量方法(汪大勋 等, 2020)、Ma 和 de la Torre (2020)提出的 GDI 和基于不完整信息矩阵(incomplete information matrix)的 Wald 检验相结合的 Stepwise 方法(为了便于理解且与本文中提出的新方法加以区分，将 Stepwise 方法称为 Wald-IC 方法)、以及 Hull 方法(Nájera et al., 2021)等。尽管一些参数化的修正方法可能存在运算量大、速度慢的不足之处，但是，这类修正方法尤其是在饱和的 CDM 框架下开发的方法的优点在于灵活性高、不需要非参数化方法那样严格的前提假设。因为饱和模型包含多类特殊模型作为特例，且在 **Q** 矩阵没有错误设定或存在少量错误时，可以较为容易地通过模型比较的方法获得恰当的特殊模型。

在饱和 CDM 框架下开发的以上 8 种参数化 **Q** 矩阵修正方法中，残差方法对于属性过度设定不敏感且在测验长度较短时统计检验力可能会偏低；当样本量较小时，TLP 方法会高估错误设定项目的数量且用于减少错误报告率的重抽样校正方法(bootstrap bagging method)的耗时可能会特别长；模拟研究表明 iGDI 的表现与 iJSD 的表现相当、甚至在一些条件下优于 iJSD (Terzi, 2017)；相对拟合统计量方法需要比较测验的所有项目关于属性所有可能组合的相对拟合值，尽管研究者提出一些减少计算次数的方法，但是在测验长度较长或属性数量较多的情况下，计算耗时仍有可能特别长。GDI 在饱和 CDM 框架下采用单个项目所有可能的属性掌握模式中正确答对概率的方差来衡量 **Q** 矩阵中相对应的 **q** 向量的区分能力，选择有最大区分能力的 **q** 向量作为正确设定的 **q** 向量。相对于 GDI 而言，iGDI 的估计效果有了一定程度的改善，但是这类方法的主要缺点是需要人为地确定一个截止值(Nájera et al., 2019)。以 GDI 研究为基础，Ma 和 de la Torre (2020)将 **Q** 矩阵修正的视角延伸到多级计分模型，在 seq-GDINA 模型(the sequential GDINA model; Ma & de la Torre, 2016)下提出了 GDI 和基于不完整信息矩阵的 Wald 检验相结合的 Wald-IC 方法。Wald-IC 方法首先采用 GDI 方法从单一属性的 **q** 向量中确定第一个所需属性，再逐步多次采用 Wald 统计量决定是否增加或删除属性来选择恰当的 **q** 向量。即，在单个项目上 Wald-IC 仅需执行 $K-1$ 个统计检验即可完成。Hull 方法试图在模型拟合与简约之间找到一种平衡以此选择恰当的 **q** 向量，研究者(Nájera et al., 2021)通过模拟研究比较了 GDI、Wald-IC 以及 Hull 方法，结果表明在大多数条件下 Hull 的表现最好、Wald-IC 的表现稍逊于 Hull。但是，Hull 和 Wald-IC 在修正错误标定的属性方面的表现较差，尤其是 **Q** 矩阵中存在较多错误设定时。研究者(Ma & de la Torre, 2020; Nájera et al., 2021)构建的 Wald-IC 统计量是使用不完整信息矩阵计算的。先前研究表明，采用不完整信息矩阵构建的统计量在后续研究中会导致一些问题，如低估模型参数标准误(Philipp et al., 2018)、用于项目功能差异检验及项目水平模型比较时导致一类错误控制率膨胀(Liu, Andersson, et al., 2019; Liu, Yin, et al., 2019; 刘彦楼 等, 2016)等。基于此，本研究认为 Wald-IC 方法在修正错误标定属性方面表现较差的主要原因可能是在 Wald 统计量的计算中采用了不完整的信息矩阵。

研究者(Liu et al., 2016; Liu, Xin, et al., 2019; Liu et al., 2021; Philipp et al., 2018; 刘彦楼 等, 2016)认为 CDM 中同时存在两种类型的模型参数：项目参数和结构参数。不完整信息矩阵(de la Torre,

2009; 2011)忽略了结构参数，计算量较小，有较大可能导致 **Q** 矩阵修正结果不够准确。以往研究者提出了多种完整信息矩阵估计方法(Liu, Xin, et al., 2019; Liu et al., 2021; Philipp et al., 2018; 刘彦楼等, 2016)，但是这些关于模型参数的信息矩阵无法直接用于 **Q** 矩阵修正中 Wald 统计量的计算，因为此类 Wald 统计量中使用的是关于模型参数的方差–协方差矩阵。此外，与其他完整信息矩阵相比，经验交叉相乘信息矩阵(empirical cross-product information matrix, XPD; Liu et al., 2021; Philipp et al., 2018; 刘彦楼 等, 2016)计算量较小，故本研究在包含全部模型参数的 XPD 矩阵的基础上，经过转换获得关于项目正确作答概率的方差–协方差矩阵，以此构建用于 **Q** 矩阵修正的 Wald 统计量(记为 Wald-XPD)。

本文的主要目的在于提出一种新的 **Q** 矩阵修正方法，并通过模拟研究与实证数据分析考察新方法的表现。模拟研究参考了以往研究者研究中采用的模拟条件(de la Torre & Chiu, 2016; Ma & de la Torre, 2020; Nájera et al., 2021)，考察新开发的方法在 **Q** 矩阵修正中的表现，并与同类方法进行比较，希望能够为实践研究者在 **Q** 矩阵修正方法的选用方面提供方法支持。本研究选择 GDI、Hull、Wald-IC 方法与 Wald-XPD 方法进行比较的原因是：首先，Wald-XPD 是在 Wald-IC 方法基础上提出的，新方法与旧方法表现的异同有待探索；其次，先前研究表明在 GDI、Hull、Wald-IC 三种方法中，Hull 的表现是最好的，故有必要比较 Hull 与 Wald-XPD 两种方法的表现；第三，限制 GDI 及 iGDI 方法实践应用的主要原因是这两种方法均需要人为地设置一个截止值，与 iGDI 相比，固定的截止值对 GDI 方法的影响相对较小(Nájera et al., 2020)，因此本研究将 GDI 也纳入比较。本文的第二部分介绍了以往研究者在饱和的 CDM 框架下提出的参数化 **Q** 矩阵修正方法。第三部分介绍了新开发的 Wald-XPD 方法。第四部分采用模拟研究，在较广泛和真实的条件下探索 Wald-XPD 方法的具体表现，并与 GDI、Hull 以及 Wald-IC 方法进行比较。第五部分探讨 Wald-XPD 方法在实证数据分析中的应用，并与 Hull 方法、Wald-IC 方法进行比较。最后对 Wald-XPD 方法进行了讨论与展望。

2 饱和 CDM 框架下的参数化 Q 矩阵修正方法

在认知诊断测验中，**Q** 矩阵是建立可观察的被试作答反应和不可观察的项目特征之间联系的桥梁。一般而言，二值计分测验中的 **Q** 是 $J \times K$ 维的矩阵，表示 J 个项目测量了 K 个属性。通常也将属性假定为二值计分，根据项目 j 是否测量了属性 k，q_{jk} 可以取 0 或者 1。假如，一份测验包含 3 个项目，共考察了 2 个属性，那么根据项目和属性之间的关系，可以构建如下 **Q** 矩阵：

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

其中，$\mathbf{q}_1' = [1,0]$ 表示测验中的第一个项目测量了属性 1(即 α_1)。但是，不同专家界定的 **Q** 矩阵不尽相同，合理设定 **Q** 矩阵并非易事。譬如，国内外研究者对于分数减法数据中(Tatsuoka, 1990)的属性设定，至今仍存在争议(de la Torre & Chiu, 2016; 汪大勋，高旭亮，蔡艳 等, 2018)。因此，对原始 **Q** 矩阵进行修正是非常必要的。

本研究以 G-DINA 模型为例，考察新提出的 Wald-XPD 方法在 **Q** 矩阵修正的表现，并与以往研究者提出的 GDI、Wald-IC、Hull 方法进行比较。G-DINA 模型是一般、饱和的 CDM 模型，对其进行适当约束，可以获得多种特殊模型(de la Torre, 2011)。令 $\boldsymbol{\alpha}_l$ 表示第 l 种属性掌握模式，$\mathbf{q}_j' = [q_{j1},\cdots, q_{jK}]$ 表示项目 j 与测验中 K 个属性之间的对应关系，在饱和的 G-DINA 模型中，正确答对项目 j 的概率可表示为：

$$p_j(\boldsymbol{\alpha}_l) = p_j(\boldsymbol{\alpha}_l, \mathbf{q}_j) = \delta_{j0} + \sum_{k=1}^{K} \delta_{jk}\alpha_{lk}q_{jk} + \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} \delta_{j,2,(k,k')}\alpha_{nk}\alpha_{nk'}q_{jk}q_{jk'} + \cdots \quad (1)$$

其中，δ_{j0} 是项目 j 的截距项参数，δ_{jk} 是 α_{lk} 的主效应参数，$\delta_{j,2,(k,k')}$ 是 α_{lk} 与 $\alpha_{lk'}$ 之间的交互效应参数。需要特别说明的是，在公式(1)中，如果 $\boldsymbol{\alpha}_l$ 或 $\mathbf{q}_j$ 中的某个元素等于 0，那么对应的项目参数 δ 也等于 0。

2.1 GDI 方法

GDI 方法(de la Torre & Chiu, 2016)是在 G-DINA 模型框架下提出的，其基本思想是：使用项目 j 中所有可能的属性掌握模式条件下的正确答对概率的方差来衡量 **q** 向量的分辨能力，选择有最大分辨能力的 **q** 向量作为正确设定的 **q** 向量，即正确设定的 **q** 向量能够使不同属性掌握模式的被试正确作答概率方差最大化。GDI 方法采用辨别指数 ς_j^2 (discriminating index)表示正确作答概率的方差，即

项目 j 的某个 **q** 向量关于所有可能的属性掌握模式的被试正确作答概率的方差：

$$\varsigma_j^2 = \sum_{l=1}^{2^K} w(\boldsymbol{\alpha}_l \mid \mathbf{x})[p_j(\boldsymbol{\alpha}_l) - \bar{p}_j]^2 \tag{2}$$

其中，$p_j(\boldsymbol{\alpha}_l)$ 表示拥有属性掌握模式为 $\boldsymbol{\alpha}_l$ 的被试正确作答的概率；$\bar{p}_j$ 表示所有被试平均的正确作答概率。另外，$w(\boldsymbol{\alpha}_l \mid \mathbf{x})$ 表示在测验项目的观察反应矩阵 **x** 中属性掌握模式为 $\boldsymbol{\alpha}_l$ 的被试的后验概率：

$$w(\boldsymbol{\alpha}_l \mid \mathbf{x}) = \sum_{i=1}^{N} w(\boldsymbol{\alpha}_l \mid \mathbf{x}_i) = \sum_{i=1}^{N} \frac{L(\mathbf{x}_i \mid \boldsymbol{\alpha}_l)\pi(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K} L(\mathbf{x}_i \mid \boldsymbol{\alpha}_l)\pi(\boldsymbol{\alpha}_l)} \tag{3}$$

公式(3)中，N 表示样本量；$L(\mathbf{x}_i \mid \boldsymbol{\alpha}_l)$ 表示属性掌握模式为 $\boldsymbol{\alpha}_l$ 的被试 i 在所有项目上作答反应 $\mathbf{x}_i$ 的条件似然函数；$\pi(\boldsymbol{\alpha}_l)$ 表示拥有第 l 种属性掌握模式的被试在总体中所占的比例，即第 l 个结构参数。

辨别指数 ς_j^2 用来衡量一个项目的辨别力，即区分不同属性掌握模式的被试的能力。有最大 GDI 且需要最少属性数的 **q** 向量，才是正确设定的 **q** 向量。但是，实践中由于随机误差，过度设定(over-specifications, OS)的 **q** 向量比正确设定的 **q** 向量有更大的 GDI 值，如全为 1 的 **q** 向量($\mathbf{q}' = [1, \cdots, 1]$)有最大的 GDI 值。因为在原有 **q** 向量的基础上增加属性会导致潜在组差异，使成功概率的方差变大，故 $\mathbf{q}' = [1, \cdots, 1]$ 时的 $\varsigma_{1:K}^2$ 是最大的。然而，这种较高的潜在组之间的差异是虚假的。本着合适与简约原则，正确设定的 **q** 向量应是简单且有最大成功作答概率方差的，故 de la Torre 和 Chiu (2016)计算了 **q** 向量的所占方差 PVAF (the proportion of variance accounted for)：

$$\text{PVAF} = \frac{\varsigma_j^2}{\varsigma_{1:K}^2} \tag{4}$$

其中，$\varsigma_{1:K}^2$ 表示项目 j 的全为 1 的 **q** 向量关于所有可能的属性掌握模式的被试正确作答概率的方差。

截止值用来判断一个 **q** 向量的 PVAF 是否合适。一个正确设定的 **q** 向量需要满足两个条件：(1)PVAF 大于截止值；(2)包含的属性数最少。若多个 **q** 向量同时满足以上两个条件，则选择 PVAF 值最大的 **q** 向量作为正确设定的 **q** 向量。

2.2　Hull 方法

Hull 方法(Nájera et al., 2021)的基本原理是：在项目水平上比较所有可能 **q** 向量的拟合指标。将所有可能的 **q** 向量呈现在 Hull 图上，Hull 图的横坐标表示与每个 **q** 向量相关的参数数量，纵坐标表示拟合指标。Hull 方法选取的拟合指标有两个：第一个是 PVAF，用来评估不同 **q** 向量的项目区分度大小；第二个是绝对模型拟合指 McFadden pseudo-R^2 (McFadden, 1974)，用于衡量观察反应中方差所占的比例，评估获得的估计值与观察反应之间的拟合度(Hull 方法的两个指标在下文分别表示为 HullP 和 HullR)。选择项目 j 中不同参数数量下有最大 PVAF 或 McFadden pseudo-R^2 值的 **q** 向量作为候选 **q** 向量，任意两个候选 **q** 向量之间会形成一条线段，将该线段下方的所有 **q** 向量移除，故 Hull 图成一条单调递增的曲线。假设项目 j 的 $K = 3$，那么以 PVAF 为指标的 Hull 图如图 1 所示，图中上方蓝色字体表示候选 **q** 向量，下方黑色字体表示该候选 **q** 向量的 PVAF。

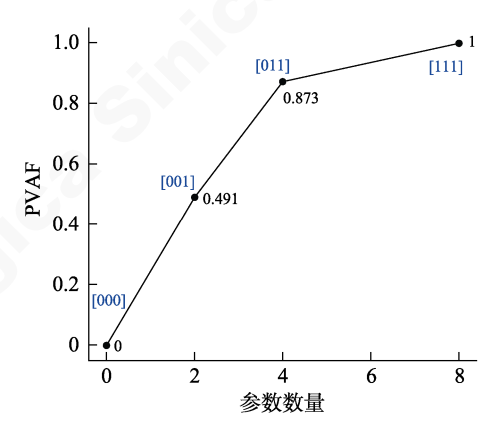


图 1　$K = 3$ 时，以 PVAF 为指标的 Hull 图

对于 Hull 方法的两个拟合指标而言，添加项目中相关联的属性会显著增加拟合指标的值；添加不关联的属性也会增加拟合指标的值，但影响可能较小。故从拟合−简约相平衡的视角出发，在 Hull 图中选择先使拟合指标显著增加，然后使拟合指标平缓增加的候选 **q** 向量作为正确设定的 **q** 向量。基于此，研究者采用 *st* 指数(Ceulemans & Kiers, 2006)计算每个候选 **q** 向量的拐角大小(the magnitude of the elbow)，选择 *st* 指数最大的候选 **q** 向量作为正确设定的 **q** 向量：

$$st_{jK} = \frac{(f_{jK} - f_{jK-1}) / (np_K - np_{K-1})}{(f_{jK+1} - f_{jK}) / (np_{K+1} - np_K)} \tag{5}$$

其中，f_{jK} 和 np_K 分别表示项目 j 的 K 个候选 **q** 向量的拟合指标和参数数量。

需要强调的是，移除候选 **q** 向量下方所有的 **q** 向量之后，若图中仅剩下原点处和全为 1 的 **q** 向量($\mathbf{q}' = [1, \cdots, 1]$)，则选择全为 1 的 **q** 向量作为该项目正确设定的 **q** 向量；若图中仍有两个或多个 **q** 向量，

则计算每个 **q** 向量的 *st* 指数，*st* 指数最大的候选 **q** 向量即为该项目正确设定的 **q** 向量。

2.3 Wald-IC 方法

用于 **Q** 矩阵修正的 Wald 统计量也是在项目水平上进行的，其基本原理是：假设项目 j 所对应的 **q** 向量定义了 2 个及以上的属性，如果将某一属性从 **q** 向量中移除而没有导致模型−数据拟合变差，那么这个属性就不是必需的。为便于理解，现举例说明。假设一个测验共测量了 2 个属性，即 $K=2$，那么，所有可能的属性掌握模式有 4 种，可以表示为：

$$\boldsymbol{\alpha}=\begin{bmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\\ \boldsymbol{\alpha}_3\\ \boldsymbol{\alpha}_4\end{bmatrix}=\begin{bmatrix}0&0\\1&0\\0&1\\1&1\end{bmatrix}$$

假设待检验的 **q** 向量为 $\mathbf{q}'=[1,1]$，那么相应的项目正确作答概率的向量可以表示为：

$$\mathbf{p}_j(\boldsymbol{\alpha})=\begin{bmatrix}p_j(\boldsymbol{\alpha}_1)\\ p_j(\boldsymbol{\alpha}_2)\\ p_j(\boldsymbol{\alpha}_3)\\ p_j(\boldsymbol{\alpha}_4)\end{bmatrix}=\begin{bmatrix}\delta_{j0}\\ \delta_{j0}+\delta_{j1}\\ \delta_{j0}+\delta_{j2}\\ \delta_{j0}+\delta_{j1}+\delta_{j2}+\delta_{j12}\end{bmatrix}$$

检验属性 1(即 α_1)是否是必需的，首先需要构建 α_1 的 **R** 矩阵：

$$\mathbf{R}=\begin{bmatrix}1&-1&0&0\\0&0&1&-1\end{bmatrix}$$

本文中，“×”用于矩阵或向量时，表示矩阵相乘。若 α_1 在统计上不是必需的，那么 $\mathbf{R}\times\mathbf{p}_j(\boldsymbol{\alpha})=\mathbf{0}$。即：

$$\begin{bmatrix}1&-1&0&0\\0&0&1&-1\end{bmatrix}\times\begin{bmatrix}p_j(\boldsymbol{\alpha}_1)\\ p_j(\boldsymbol{\alpha}_2)\\ p_j(\boldsymbol{\alpha}_3)\\ p_j(\boldsymbol{\alpha}_4)\end{bmatrix}=\begin{bmatrix}-\delta_{j1}\\ -(\delta_{j1}+\delta_{j12})\end{bmatrix}=\mathbf{0}$$

表明掌握 α_1 不会增加正确答对项目 j 的概率，故 α_1 不是必需的。此外，检验属性 2 (即 α_2)是否必需的 **R** 矩阵为：

$$\mathbf{R}=\begin{bmatrix}1&0&-1&0\\0&1&0&-1\end{bmatrix}$$

需要说明的是，对于项目 j 而言，不同的待检验属性所对应的 **q** 向量是不同，也就是项目参数估计值是不同的，因此，向量 $\mathbf{p}_j(\boldsymbol{\alpha})$ 的值不是固定的。

Wald-IC 统计量的形式为：

$$\text{Wald}_{(\text{IC})}=[\mathbf{R}\times\mathbf{p}_j(\boldsymbol{\alpha})]'(\mathbf{R}\times\mathbf{V}_{(\text{IC})j}\times\mathbf{R}')^{-1}[\mathbf{R}\times\mathbf{p}_j(\boldsymbol{\alpha})] \quad (6)$$

其中，$\mathbf{V}_{(\text{IC})j}$ 是基于不完整信息矩阵计算的项目 j 正确作答概率的方差−协方差矩阵。

Wald-IC 方法修正 **Q** 矩阵的步骤为：首先，需要构建一个 $2^{K_j^*-1}\times 2^{K_j^*}$ 的 **R** 矩阵，K_j^* 表示待检验的 **q** 向量中定义的项目 j 需要的属性数量。在零假设下，即属性 k 在统计上不是必需的，那么 $\mathbf{R}\times\mathbf{p}_j(\boldsymbol{\alpha})=\mathbf{0}$。其次，需要对不完整信息矩阵求逆获得项目正确作答概率的方差−协方差矩阵 $\mathbf{V}_{(\text{IC})j}$ 来构建 Wald 统计量。Ma 和 de la Torre (2020)采用的是 de la Torre (2009)提出的考虑全部项目正确作答概率的不完整信息矩阵估计方法 $\mathcal{I}_{D09}$：

$$\mathcal{I}_{D09}=\frac{\partial\ell(\mathbf{x})}{\partial[\mathbf{p}_j(\boldsymbol{\alpha})]}\times\frac{\partial\ell(\mathbf{x})}{\partial[\mathbf{p}_j(\boldsymbol{\alpha})]'} \quad (7)$$

其中，$\ell(\mathbf{x})$ 表示观察数据的对数似然函数。理论上，用于 **Q** 矩阵修正的 Wald 统计量渐近 χ^2 分布，自由度是 $2^{K_j^*-1}$。但是，Wald-IC 统计量中方差−协方差矩阵的计算存在不准确的问题，可能导致 **Q** 矩阵修正的效果不理想。

3 基于完整 XPD 矩阵的 Wald-XPD 方法

3.1 使用 XPD 矩阵构建 Wald-XPD 统计量

Philipp 等人(2018)和 Liu 等人(2021)用结构参数 $\boldsymbol{\pi}$ 描述被试总体的潜在属性掌握模式 $\boldsymbol{\alpha}$ 的分布状况，$\boldsymbol{\pi}=(\pi_1,\cdots,\pi_L)'$。假设 $K=2$，那么在这个测验中被试所有可能的属性掌握模式 $\boldsymbol{\alpha}_l$ 有 $L=4$ 种，$\pi_l=\pi(\boldsymbol{\alpha}_l)$ 表示被试总体中具有第 l 种属性掌握模式 $\boldsymbol{\alpha}_l$ 的分布比例。例如，$\pi(\boldsymbol{\alpha}_1)$ 是被试总体中具有第 1 种属性掌握模式 $\boldsymbol{\alpha}_1'=[0,0]$ 的分布比例。

研究者提出了很多完整信息矩阵的估计方法，主要有：完整的经验交叉相乘信息矩阵(Liu et al., 2021; Philipp et al., 2018; 刘彦楼 等, 2016)、完整的观察信息矩阵(observed information matrix; Liu et al., 2021; 刘彦楼 等, 2016)、完整的三明治信息矩阵(sandwich-type information matrix; Liu, Xin, et al., 2019; Liu et al., 2021)等。由于考虑所有模型参数，完整信息矩阵的计算量较大，尤其是观察信息矩阵以及三明治信息矩阵涉及观察数据的对数似然函数关于所有模型参数的二阶偏导，计算量非常大。本文采用观察数据对数似然函数关于项目参数 $\boldsymbol{\delta}$ 和结构参数 $\boldsymbol{\pi}$ 的一阶导向量交叉相乘而计算的 XPD 矩阵：

$$\mathcal{I}_{\text{XPD}}=\begin{bmatrix}\frac{\partial\ell(\mathbf{x})}{\partial\delta_1}\times\frac{\partial\ell(\mathbf{x})}{\partial\delta_1} & \cdots & \frac{\partial\ell(\mathbf{x})}{\partial\delta_1}\times\frac{\partial\ell(\mathbf{x})}{\partial\pi_{L-1}}\\ \vdots & \ddots & \vdots\\ \frac{\partial\ell(\mathbf{x})}{\partial\pi_{L-1}}\times\frac{\partial\ell(\mathbf{x})}{\partial\delta_1} & \cdots & \frac{\partial\ell(\mathbf{x})}{\partial\pi_{L-1}}\times\frac{\partial\ell(\mathbf{x})}{\partial\pi_{L-1}}\end{bmatrix} \quad (8)$$

在构建 Wald 统计量之前，本研究首先对 XPD 矩阵做了以下三个方面的处理：

(1)对 XPD 矩阵求逆，获得方差−协方差矩阵 $\mathbf{\Sigma}_{\mathrm{XPD}}$，即：$\mathbf{\Sigma}_{\mathrm{XPD}} = \mathcal{I}_{\mathrm{XPD}}^{-1}$。选取 $\mathbf{\Sigma}_{\mathrm{XPD}}$ 中项目 j 对应方差−协方差矩阵 $\mathbf{\Sigma}_j$。

(2)采用 $\mathbf{M}_j$ 矩阵(de la Torre, 2011)通过矩阵乘法将项目参数的方差−协方差矩阵 $\mathbf{\Sigma}_j$，转换为项目正确作答概率的方差−协方差矩阵 $\mathbf{V}_{(\mathrm{XPD})j}$，即：$\mathbf{V}_{(\mathrm{XPD})j} = \mathbf{M}_j \times \mathbf{\Sigma}_j$。$\mathbf{M}_j$ 矩阵是 $2^{K_j^*} \times 2^{K_j^*}$ 维的矩阵，表示项目 j 中各个属性掌握模式与项目参数之间的对应关系，可以将项目参数转换为各个属性掌握模式下的正确作答概率。例如，假设项目 j 中 K_j^*=2，则对于饱和 G-DINA 模型而言该项目的 $\mathbf{M}_j$ 矩阵可以表示为：

$$\mathbf{M}_{j[4\times4]} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

通过 $\mathbf{M}_j$ 矩阵，可以获得该项目中各个属性掌握模式下的正确作答概率向量 $\mathbf{p}_j(\boldsymbol{\alpha})$：

$$\mathbf{p}_j(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \delta_{j0} \\ \delta_{j1} \\ \delta_{j2} \\ \delta_{j12} \end{bmatrix} = \begin{bmatrix} \delta_{j0} \\ \delta_{j0} + \delta_{j1} \\ \delta_{j0} + \delta_{j2} \\ \delta_{j0} + \delta_{j1} + \delta_{j2} + \delta_{j12} \end{bmatrix}$$

根据统计学中模型参数方差−协方差矩阵的性质(或参考 Li & Wang, 2015)，可以通过 $\mathbf{M}_j$ 矩阵将项目参数的方差−协方差矩阵 $\mathbf{\Sigma}_j$ 转换为项目正确作答概率的方差−协方差矩阵 $\mathbf{V}_{(\mathrm{XPD})j}$。因此，基于 XPD 矩阵构建 Wald 统计量的形式为：

$$\mathrm{Wald}_{(\mathrm{XPD})} = [\mathbf{R} \times \mathbf{p}_j(\boldsymbol{\alpha})]'(\mathbf{R} \times \mathbf{V}_{(\mathrm{XPD})j} \times \mathbf{R}')^{-1}[\mathbf{R} \times \mathbf{p}_j(\boldsymbol{\alpha})] \tag{9}$$

(3)对比完整和不完整信息矩阵可知，完整信息矩阵考虑模型中的全部参数，计算量较大，修正过程较为耗时。故本研究采用 C++语言编写 XPD 矩阵，提高 **Q** 矩阵修正的速度。

3.2　Wald-XPD 方法的具体实施步骤

Wald-XPD 方法用于 **Q** 矩阵修正是逐个项目进行的。假设项目 j 的 **q** 向量的集合是由单一属性构成的。**A** 是所需属性的集合，**B** 是需要修正的目标属性的集合，修正之初，$\mathbf{A} = \varnothing$，$\mathbf{B} = \{1, 2, \cdots, K\}$。

本研究新提出的 Wald-XPD 方法的修正步骤如下：

步骤(1)：选择项目 j 中具有最大 PVAF 值的单一属性 **q** 向量中包含的属性为第一个所需属性，更新集合 **A**、**B**。

步骤(2)：将该单一属性 **q** 向量的 PVAF 值与 0.95 进行比较，大于 0.95 说明该 **q** 向量是合适的，停止修正，否则继续修正。

步骤(3)：更新集合 **A**、**B**。选出具有较大 PVAF 的 **q** 向量进行修正，将该 **q** 向量中各属性使用 Wald-XPD 统计量进行显著性检验，确定该 **q** 向量对应的集合 **A** 和集合 **B** 中的属性是否应该移除或添加，然后判断 **q** 向量的 PVAF 是否大于 0.95，大于 0.95 说明这个 **q** 向量是合适的，停止修正，否则继续修正。

步骤(4)：重复步骤(3)，直到某个 **q** 向量的 PAVF 值大于 0.95，或者没有属性移除或添加则停止修正。

步骤(5)：在单个项目修正结束后，重新计算 PVAF 以及 Wald-XPD 统计量，直到达到最大迭代或者某次迭代结束后的 **q** 向量与前一次迭代的 **q** 向量完全相等则停止修正。

为了便于理解，现举例说明 Wald-XPD 方法用于某个项目的 **q** 向量的修正算法。假设项目 j 中 **q** 向量的属性数 $K = 3$，Wald-XPD 方法修正该 **q** 向量的过程如图 2 所示。

4　模拟研究

模拟研究的目的是在较为广泛和真实的条件下探讨 Wald-XPD 方法在 **Q** 矩阵修正中的表现，并与以往研究者提出的 GDI、Wald-IC 以及 Hull (HullP、HullR)方法进行比较。

4.1　方法

4.1.1　研究设计

为便于比较，本研究参考以往研究设计(Ma & de la Torre, 2020; Nájera et al., 2021)，共操纵了 5 种因素：项目数和属性数的比例(ratio of number of items to attribute, *JK*)、样本量(*N*)、**Q** 矩阵错误设定的比例(Q-matrix misspecification rate, *QM*)、属性分布(attribute distribution, *AD*)、项目质量(item quality, *IQ*)。本研究将属性数设置为 $K = 4$，因为这是应用类文章中最经常出现的属性数(Nájera et al., 2020)。以往研究中常用的项目数是 11 到 30 (Sessoms & Henson, 2018)，故本研究将项目数设置为 16 和 32，所以，本研究共考虑 2 种测验结构：$J = 16[(K = 4) \times (JK = 4)]$、$J = 32[(K = 4) \times (JK = 8)]$。样本量有两个

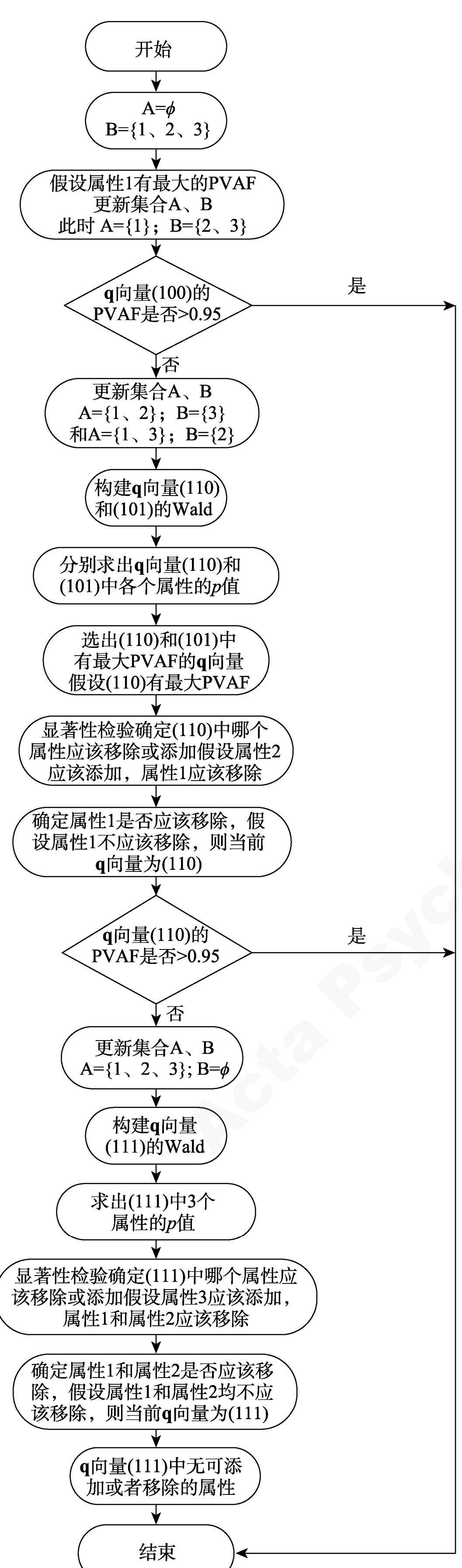


图 2　Wald-XPD 方法用于 $\mathbf{q}_j$ 向量的修正流程图

水平：500 和 1000 (Chen, 2017; de la Torre, 2011; Ma & de la Torre, 2016)，分别代表小样本和大样本。本研究共有 48 个实验条件，各因素水平如表 1 所示。

表 1　模拟研究中各因素水平汇总

因素	因素水平
样本量 N	500、1000
项目数和属性数的比例 JK	4、8
属性数 K	4
平均项目质量 IQ	0.4、0.6、0.8
属性分布 AD	均匀分布、高阶分布
错误设定的比例 QM	0.15、0.3
链接函数	G-DINA 模型
Q 矩阵修正方法	GDI、Wald-IC、Hull (HullP、HullR)、Wald-XPD

4.1.2　数据生成

被试的属性掌握模式从两种分布中产生：均匀分布和高阶分布(de la Torre & Douglas, 2004)。对于均匀分布，每个被试的属性掌握模式是从所有可能的属性掌握模式中以相等的概率随机生成的；对于高阶分布，被试的能力(θ)来自于标准正态分布，属性难度参数 δ_k 在[–1.5, 1.5]之间给出等距值(Ma & de la Torre, 2020)。

项目质量分为高、中、低 3 个水平。高项目质量：$P_j(0) \sim U(0, 0.2)$且 $P_j(1) \sim U(0.8, 1)$；中等项目质量：$P_j(0) \sim U(0.1, 0.3)$且 $P_j(1) \sim U(0.7, 0.9)$；低项目质量：$P_j(0) \sim U(0.2, 0.4)$且 $P_j(1) \sim U(0.6, 0.8)$。其中，$P_j(0)$表示仅凭猜测答对的概率，$P_j(1)$表示掌握项目所要求的全部属性的被试答对该项目的概率。成功的概率有两种限制：(1)项目反应函数在属性数上具有单调性；(2)与单个属性相联系的项目参数的总和限制为大于 0.15。这两个条件保证所有的属性都具有不可忽视的作用。

真实 **Q** 矩阵符合以下限制：(1)每个 **Q** 矩阵至少包含两个单位矩阵(identity matrix)；(2)除了两个单位矩阵外，每个项目至少测量一个属性；(3) **Q** 矩阵由 1 个属性 **q** 向量(50%)、2 个属性 **q** 向量(25%)和 3 个属性 **q** 向量(25%)组成。这个比例主要是参考之前研究(Nájera et al., 2021)，使用较高比例的单一属性 **q** 向量的原因是满足每个 **Q** 矩阵至少包含两个单位矩阵的模型可识别条件(Gu et al., 2018)。错误设定的 **Q** 矩阵的比例为：0.15 和 0.3。错误设定是在两个约束条件下随机引入：(1)所有项目必须至少测量一个属性；(2)始终保留一个单位矩阵。

在每个条件下，均生成 500 个数据集，每个数

据集中生成新的真实 **Q** 矩阵和项目参数。所有的模拟研究和分析都在 R 软件中进行。

4.1.3　评价指标

QRR (Q-matrix recovery rate)用来测量 **Q** 矩阵的恢复比例, 可以表示为:

$$\mathrm{QRR}=\frac{\sum_{J=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(s)}=q_{jk}^{(t)}\right)}{J\times K} \tag{10}$$

其中, $I(\bullet)$ 是指示函数, 若修正前后项目 j 的 **q** 向量完全一致, 则 $I\left(q_{jk}^{(s)}=q_{jk}^{(t)}\right)=1$, 否则 $I\left(q_{jk}^{(s)}=q_{jk}^{(t)}\right)=0$ 。$q_{jk}^{(s)}$ 和 $q_{jk}^{(t)}$ 分别表示项目 j 中属性 k 的建议 q 元素和真实 q 元素。

TPR (true positive rate)表示保留正确标定属性的比例:

$$\mathrm{TPR}=\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(s)}=q_{jk}^{(t)}\mid q_{jk}^{(o)}=q_{jk}^{(t)}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(o)}=q_{jk}^{(t)}\right)} \tag{11}$$

其中, $q_{jk}^{(o)}$ 表示项目 j 中属性 k 的原始 q 元素。

TNR (true negative rate)表示修正错误标定属性的比例:

$$\mathrm{TNR}=\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(s)}=q_{jk}^{(t)}\mid q_{jk}^{(o)}\neq q_{jk}^{(t)}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(o)}\neq q_{jk}^{(t)}\right)} \tag{12}$$

本研究除了使用 QRR、TPR、TNR 来考察各个方法总体的表现之外, 还参考其他指标来获得更加全面具体的结果。OS 表示过度设定, US (under-specifications)表示吝啬设定, 表达式分别为:

$$\mathrm{OS}=\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(s)}>q_{jk}^{(t)}\right) \tag{13}$$

$$\mathrm{US}=\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{(s)}<q_{jk}^{(t)}\right) \tag{14}$$

以上 5 个指标从不同方面反映了 **Q** 矩阵的修正效果。其中, QRR、TPR、TNR 的值越高, 表示该修正方法的 **Q** 矩阵恢复率以及保留正确标定属性和修正错误标定属性的比例越高, 修正效果越好。OS 和 US 的值越小, 表示该修正方法存在较少过度设定和吝啬设定的趋势, 修正效果越好。

4.2　研究结果

4.2.1　GDI、Hull、Wald-IC 以及 Wald-XPD 在各因素不同水平上的表现

表 2 呈现了 GDI、Hull (HullP、HullR)、Wald-IC 以及 Wald-XPD 方法在各因素不同水平上的 QRR、TPR、TNR、OS 和 US 值, 表中加粗数据是相同条件下的最优结果。

首先, 比较的是各实验条件的综合影响。**Q** 矩阵错误设定的比例、项目质量、样本量以及属性分布对于 GDI、Wald-IC、Hull (HullP、HullR)以及 Wald-XPD 方法在各个指标上的表现有明显影响。除 Hull (HullP、HullR)方法的 TPR 指标受项目质量的影响较小外, 在项目质量较高的条件下, 所有方法的表现均优于其他水平。**Q** 矩阵错误设定的比例和样本量对于 4 种方法在各个指标上的表现也存在一定的影响, 随着 **Q** 矩阵错误设定的比例降低和样本量增大, 4 种方法均有更好的 **Q** 矩阵修正表现。均匀分布下, 4 种方法在各个指标上的表现均优于高阶分布。就 JK 因素而言, JK 对于 GDI、Wald-IC 和 Wald-XPD 在 QRR 指标上的表现, 以及所有的修正方法在 TNR 指标上的表现影响明显, 所有指标在 $JK=8$ 水平下的结果优于 $JK=4$。

其次, 比较的是 4 种修正方法的综合表现。所有方法在 QRR 以及 TPR 指标上没有表现出明显优劣。其中, 本研究中新提出的 Wald-XPD 在 TNR 指标上的表现明显优于其他方法; GDI 在 OS 指标上的表现较优, 但是在 US 指标上表现相对较差; HullR 在 OS 指标上的表现较差, 但是在 US 指标上表现相对较优; Wald-IC 在 US 指标上表现相对较差。

根据以上综合比较可知, Wald-XPD 以及 HullP 在各个指标上有相对较好的表现, 且在 TNR 指标上 Wald-XPD 的表现最好。此外, 鉴于 Wald-XPD 是在 Wald-IC 基础上新提出的方法, 故接下来本研究主要探讨 Wald-XPD、Wald-IC 以及 HullP 方法在 QRR、TPR 以及 TNR 这 3 个主要指标上的具体表现, 并重点关注 Wald-XPD 在 TNR 指标上的表现, 即 Wald-XPD 修正 **Q** 矩阵中错误标定属性的能力。

4.2.2　Wald-XPD 在修正错误标定属性时的表现

图 3 呈现的是 HullP、Wald-IC 以及 Wald-XPD 方法在 48 种具体的模拟条件下获得的 QRR 的值。由图 3 可知, 项目质量对于这 3 种方法的表现影响最为明显, 随着项目质量的提高, QRR 的值也在增加。另外, 样本量、**Q** 矩阵错误设定的比例以及属性分布对于这 3 个方法在 QRR 指标上的表现稍有影响, 且趋势一致。就 QRR 指标而言, HullP、Wald-IC 以及 Wald-XPD 方法的表现仅有细微差异, 即当 $IQ=0.4$ 时 Wald-XPD 的表现略微低于另外两种方法。

图 4 呈现的是 3 种方法在 TPR 指标上的表现。由图 4 可知, 在所有条件下 Wald-IC 以及 HullP 方

表 2 不同因素水平的结果

指标	方法	*QM*		*IQ*			*N*		*JK*		*AD*	
		0.15	0.3	0.4	0.6	0.8	500	1000	4	8	均匀分布	高阶分布
QRR	GDI	0.906	0.828	0.859	0.922	0.945	0.922	0.922	0.906	0.930	0.938	0.906
	Wald-IC	**0.945**	0.813	0.844	0.922	**0.969**	0.906	0.938	0.891	0.930	0.938	0.906
	HullP	0.930	0.852	**0.875**	**0.945**	0.953	**0.938**	**0.953**	**0.938**	**0.945**	**0.953**	**0.930**
	HullR	0.891	0.797	0.844	0.891	0.922	0.898	0.906	0.906	0.906	0.914	0.891
	Wald-XPD	0.937	**0.867**	0.820	0.938	**0.969**	0.906	**0.953**	0.906	**0.945**	**0.953**	0.906
TPR	GDI	0.944	0.922	0.933	0.936	0.953	0.936	0.945	0.944	0.936	0.954	0.926
	Wald-IC	0.945	0.933	0.908	0.954	**0.969**	0.933	0.956	0.944	0.945	0.956	0.938
	HullP	**0.963**	**0.936**	**0.963**	**0.961**	0.956	**0.953**	**0.969**	**0.963**	**0.956**	**0.967**	**0.953**
	HullR	0.936	0.911	0.953	0.927	0.930	0.927	0.944	0.956	0.922	0.944	0.926
	Wald-XPD	0.944	0.900	0.835	0.944	**0.969**	0.917	0.953	0.920	0.944	0.953	0.927
TNR	GDI	0.800	0.684	0.421	0.789	0.900	0.711	0.737	0.579	0.842	0.800	0.684
	Wald-IC	0.789	0.579	0.405	0.700	0.900	0.632	0.684	0.526	0.789	0.700	0.632
	HullP	0.800	0.684	0.368	0.833	**0.947**	0.737	0.800	0.600	0.895	0.816	0.700
	HullR	0.684	0.579	0.263	0.676	0.895	0.600	0.632	0.421	0.763	0.684	0.579
	Wald-XPD	**0.900**	**0.816**	**0.684**	**0.900**	**0.947**	**0.840**	**0.894**	**0.700**	**0.920**	**0.900**	**0.830**
OS	GDI	**0**	**3**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	Wald-IC	1	5	**3**	1	**0**	1	**0**	1	**0**	**0**	1
	HullP	1	5	5	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	HullR	8	11	9	9	6	8	8	5	11	7	8
	Wald-XPD	1	**3**	4	1	**0**	2	1	1	1	1	1
US	GDI	7	10	9	7	5	7	6	5	9	5	8
	Wald-IC	6	10	11	6	3	8	5	5	8	5	7
	HullP	5	8	6	5	4	5	4	3	6	4	6
	HullR	**2**	**5**	**5**	**2**	**1**	**2**	**1**	**1**	**2**	**1**	2
	Wald-XPD	5	8	12	4	3	7	5	5	6	4	7

注：粗体表示各指标不同水平下的最好结果。

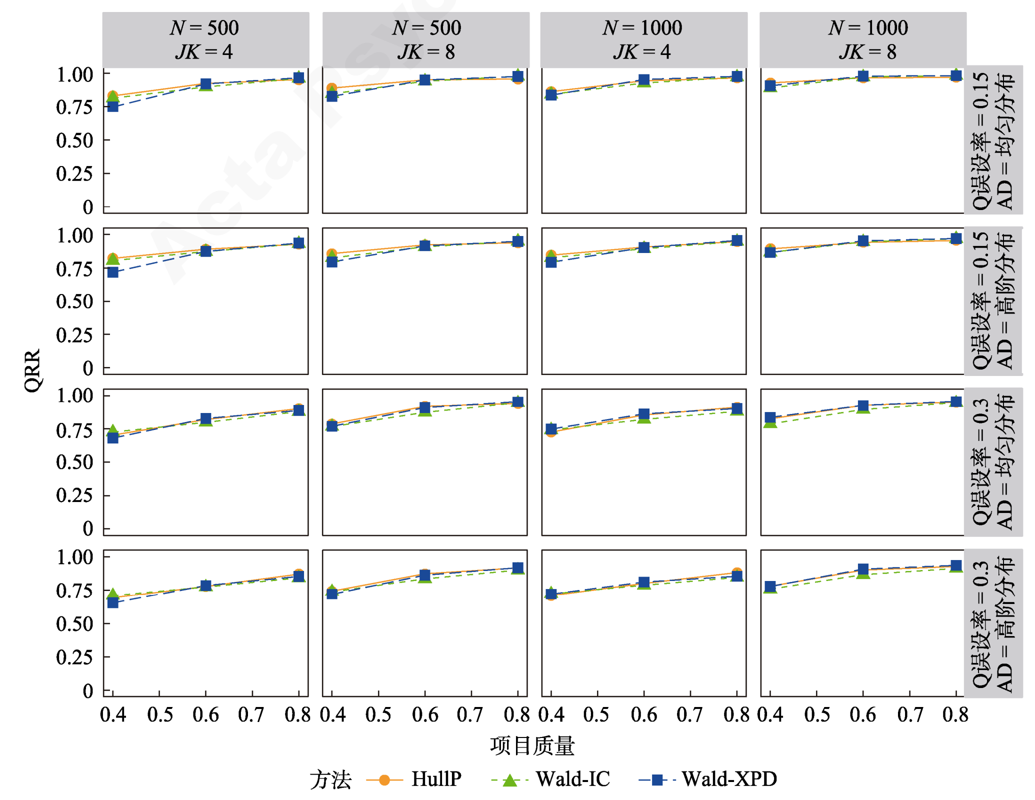


图 3 HullP、Wald-IC 与 Wald-XPD 方法在 QRR 指标上的表现

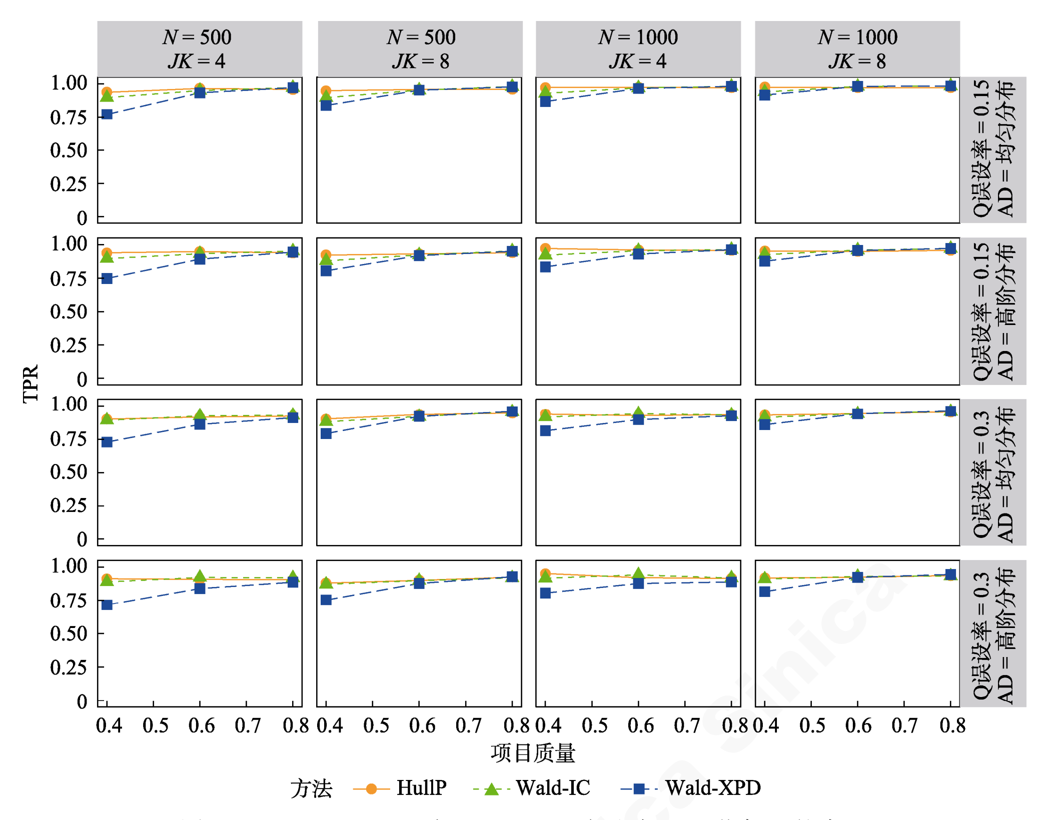


图 4 HullP、Wald-IC 与 Wald-XPD 方法在 TPR 指标上的表现

法均能获得较高的 TPR 值。项目质量对于 Wald-XPD 方法的表现有一定的影响，当项目质量较低时，Wald-XPD 在 TPR 指标上的表现不如 Wald-IC 以及 HullP 方法；随着项目质量的提高，3 种方法在 TPR 指标上的表现相当。

图 5 呈现的是 3 种方法在 TNR 指标上的表现。

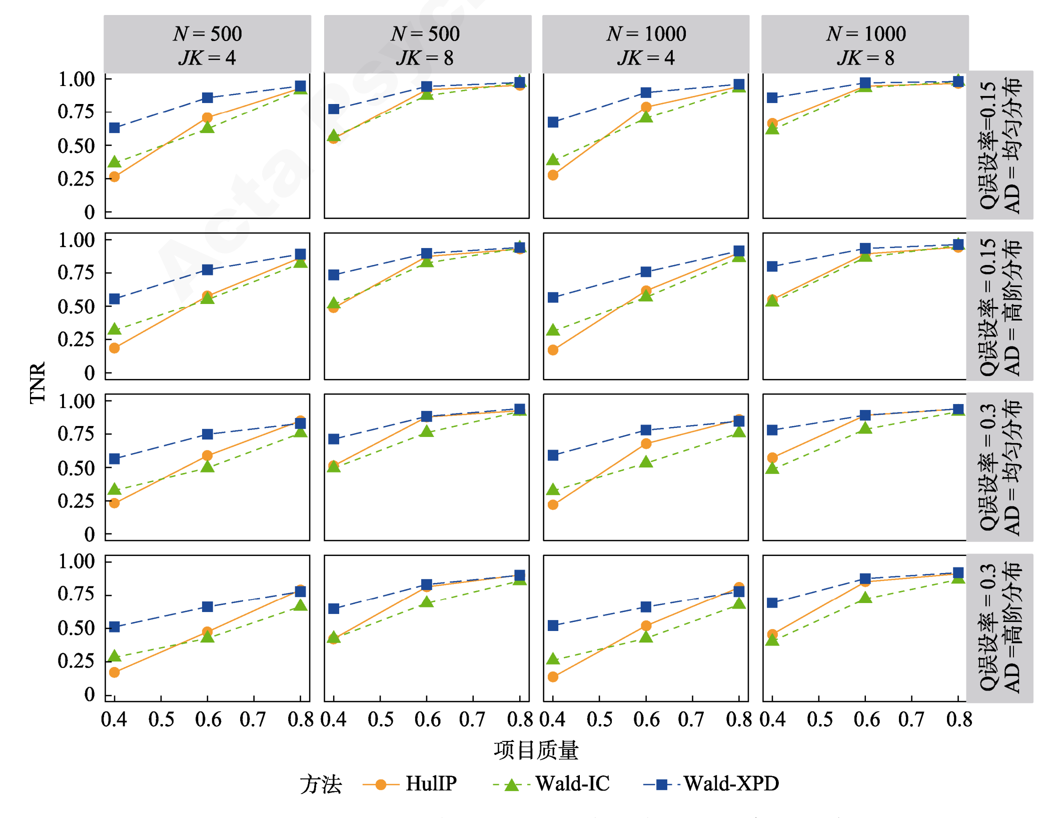


图 5 HullP、Wald-IC 与 Wald-XPD 方法在 TNR 指标上的表现

在所有条件下，Wald-XPD 方法在 TNR 指标上的表现均是最优的，对比 Wald-XPD 方法在 TPR 及 TNR 上的表现可知，低项目质量条件对这个方法产生了一些不利影响，而在中等或高项目质量条件下，Wald-XPD 能有效保留 **Q** 矩阵中正确标定的属性，也能有效修正 **Q** 矩阵中错误标定的属性。测验长度较短、项目质量较低及 **Q** 矩阵错误设定比例较高时 HullP 方法的表现较差，结合同样条件下 HullP 在 TPR 指标上的表现可知，虽然 HullP 方法在保留正确标定属性方面略微优于 Wald-XPD，但是它较多地保留了错误标定的属性。即，HullP 方法倾向于较少地修正原始 **Q** 矩阵中的属性。在低项目质量条件下的多数情景中，虽然 Wald-IC 方法在 TNR 上的表现优于 HullP，但是在随着项目质量的提高 HullP 在多数情景中的表现优于 Wald-IC。HullP、Wald-IC 以及 Wald-XPD 方法在 TNR 指标上的表现受样本量、测验长度、项目质量、属性分布及错误设定比例的影响明显。随着 **Q** 矩阵错误设定比例降低、项目质量提高、测验长度增加，HullP 和 Wald-IC 方法的 TNR 值有所提高，但仍低于 Wald-XPD 方法的 TNR 值。

5　实证数据分析

本研究采用实证数据进一步考察 Wald-XPD 方法的表现，并与 HullP、Wald-IC 方法进行比较。被试反应数据及测验项目获取自 R 软件包 *pks* (Heller & Wickelmaier, 2013)，来自德国图宾根(Tuebingen)大学的一个学习实验，包含 504 名被试在 12 个概率论测验项目上的作答。Philipp 等人(2018)认为这个数据集共测试了 4 种不同的属性：α_1(计算某事件发生概率)、α_2(计算某事件的对立事件发生的概率)、α_3(计算两个无关事件同时发生的概率)、α_4(计算两个独立事件发生的概率)，并定义了表 3 所示的原始 **Q** 矩阵。

本研究在饱和 G-DINA 模型框架下，使用 HullP、Wald-IC 以及 Wald-XPD 方法对原始 **Q** 矩阵进行了修正。表 3 中的结果显示，HullP 方法共修正了 6 个元素，Wald-IC 方法共修正了 5 个元素，Wald-XPD 方法一共修正了 16 个元素，Wald-IC 方法修正的 5 个元素均包括在 Wald-XPD 方法修正的元素之中。使用相对拟合、绝对拟合及近似拟合指标比较原始 **Q** 矩阵、HullP、Wald-IC 及 Wald-XPD 方法修正后的 **Q** 矩阵的模型−数据拟合表现。拟合指标包括：相对拟合指标 AIC (Akaike information criterion)和 BIC (Bayesian information criterion)、有限信息绝对拟合(limited-information absolute fit)指标 M_2 及近似拟合指标 RMSEA$_2$ (root mean square error of approximation; Liu et al., 2016)，结果见表 4。就相对拟合指标而言，$\mathbf{Q}_{\text{HullP}}$ 获得最佳的 AIC 指标，$\mathbf{Q}_{\text{XPD}}$ 的 AIC 指标与其接近；$\mathbf{Q}_{\text{XPD}}$ 获得最佳的 BIC 指标，其次是 $\mathbf{Q}_{\text{IC}}$，$\mathbf{Q}_{\text{HullP}}$ 的 BIC 指标最差。即，Wald-XPD 方法修正后的 **Q** 矩阵的相对拟合指标更

表 3　原始 Q 矩阵以及各方法对属性的修正情况

项目	原始 Q 矩阵			
	α_1	α_2	α_3	α_4
1	1	0	0	0
2	0	1*	0*	0
3	0	0	1	0
4	0	0	0	1
5	1*	1	0	0^
6	1*	1	0	0
7	1*	0*	1*	0
8	1*	0*	1	0*
9	1	0	0	1*#^
10	0	1*#^	0	1
11	1*#^	1*#^	0	1
12	1*	0	1*#^	1

注：*为 Wald-XPD 方法调整的属性，#为 Wald-IC 方法调整的属性，^为 HullP 方法调整的属性

表 4　基于 3 种方法修正前后 Q 矩阵的拟合指标

Q	相对拟合指标		有限信息拟合指标			
			M_2			RMSEA$_2$
	AIC	BIC	M_2	*df*	*p*	
$\mathbf{Q}_{\text{original}}$	4979.256	5245.278	23.919	15	0.067	0.0343
$\mathbf{Q}_{\text{XPD}}$	4962.484	5152.500	51.991	33	0.019	0.0338
$\mathbf{Q}_{\text{IC}}$	4964.200	5171.110	50.051	29	0.009	0.0380
$\mathbf{Q}_{\text{HullP}}$	4954.912	5178.709	40.037	25	0.029	0.0345

优。在绝对拟合指标 M_2 上，$\mathbf{Q}_{\text{IC}}$ 的 $p < 0.01$，表明 Wald-IC 方法修正的 **Q** 矩阵与数据失拟；$\mathbf{Q}_{\text{HullP}}$ 和 $\mathbf{Q}_{\text{XPD}}$ 的 p 值分别为：0.029 和 0.019，表明 HullP 和 Wald-XPD 方法修正后的 **Q** 矩阵没有在 0.01 显著性水平上拒绝模型−数据拟合的原假设。对于 RMSEA_2 指标而言，其值越接近 0 修正效果越好，其中 $\mathbf{Q}_{\text{XPD}}$ 的 RMSEA_2 最接近于 0，即 $\mathbf{Q}_{\text{XPD}}$ 在 RMSEA_2 指标上有最好的表现(Liu et al., 2016)。综合考虑相对拟合、绝对拟合和近似拟合指标，本研究认为 Wald-XPD 方法修正后的 **Q** 矩阵在模型−数据拟合方面表现最优。

需要特别说明的是，本研究的目的是在一般性的 CDM 框架下开发具有广泛适用性的 **Q** 矩阵修正方法。因此，实证数据分析的重点是原始 **Q** 矩阵的修正，没有在饱和 G-DINA 模型的基础上进一步在项目水平上进行模型比较(Liu, Andersson, et al., 2019)。另外，M_2 统计量在模型参数过度设定时，即模型中冗余参数过多时，可能存在统计检验力不足的问题(参考 Chen et al., 2018)。举例而言，对比原始 $\mathbf{Q}_{\text{original}}$ 矩阵及修正后的 $\mathbf{Q}_{\text{XPD}}$ 矩阵可知，$\mathbf{Q}_{\text{original}}$ 中可能存在较多过度设定的元素，因此，导致 $\mathbf{Q}_{\text{original}}$ 的 M_2 统计量的 p 值大于 0.01。参考先前研究(Liu et al., 2016)，本文认为在模型−数据拟合评价方面，近似拟合统计量 RMSEA_2 可能更具参考价值。

根据表 3 的结果可知, Wald-XPD 方法修正的属性中，对 α_1 修正最多，共修正 6 个题目，均是将 α_1 从 1 变成 0。例如，第 6 题“一个盒子包含 20 个以下颜色的大理石：4 个白色，14 个绿色，2 个红色。随机抽取的大理石不是白色的概率是多少？”解决这个问题可以先计算出该事件的对立事件发生的概率 (α_2)，即随机抽取的大理石是白色的概率，然后再用 1 减去该对立事件发生的概率即可得出正确结论。对于 5、6、7 题来说，当被试掌握 α_2 时即能够解决问题，故 α_1 不是必需的。再如，第 11 题“车库里有 50 辆车。20 辆是黑色的，10 辆是柴油动力的。假设汽车的颜色与燃料种类无关。随机选择的汽车不是黑色的，而是柴油动力的概率是多少？”题中汽车颜色与燃料种类是独立事件，计算随机选择的汽车不是黑色的而是柴油动力的概率即两个独立事件发生的概率 (α_4)，当被试掌握 α_4 时即能够解决问题，故 α_1 不是必需的。在 5、6、7、8、11、12 这 6 道题中，α_1 不是必需的，Wald-XPD 方法均正确修正了错误标定的 α_1。所以说，使用 Wald-XPD 修正方法获得的 $\mathbf{Q}_{\text{XPD}}$ 矩阵在理论上具有合理性。

值得注意的是，本研究中提出的 **Q** 矩阵修正方法是从作答数据出发的，在一定程度上可以避免专家标定 **Q** 矩阵的主观性，减轻专家负担，但是客观方法标定的 **Q** 矩阵不能直接作为最终的 **Q** 矩阵，应该作为专家标定 **Q** 矩阵的重要参考(Xu & Shang, 2018)。

6 讨论与展望

6.1 结论与讨论

CDM 依赖正确设定的 **Q** 矩阵以获得准确的属性剖面分类(Rupp & Templin, 2008)。以往研究者提出的 GDI、Wald-IC、Hull 方法在多数的应用情景中虽然有较好的表现，但这些方法对 **Q** 矩阵中错误标定的属性不够敏感。本研究提出使用完整的 XPD 矩阵计算用于 **Q** 矩阵修正的方法(Wald-XPD 方法)，并系统探讨了样本量、测验长度、**Q** 矩阵错误设定比例、属性分布等因素对 **Q** 矩阵修正结果的影响。采用实证数据展示了新提出的 Wald-XPD 方法在实际应用中的表现与价值。

本研究结果表明：(1)整体而言，Wald-XPD 方法的表现优于 GDI、Hull、Wald-IC 方法。Wald-XPD 方法能够弥补 GDI、Hull、Wald-IC 方法在一些条件下对于错误标定属性不敏感的不足之处，且在 **Q** 矩阵恢复率和保留正确标定属性的比例方面也有较好的表现。(2) GDI、Hull、Wald-IC 和 Wald-XPD 方法随着项目质量的提高、样本量增大、测验长度增加以及 **Q** 矩阵错误设定比例的降低，在修正 **Q** 矩阵上有更好的表现。(3)由 HullP、Wald-IC 以及 Wald-XPD 方法进一步比较的结果可知，3 种方法在 **Q** 矩阵恢复率方面差异较小，HullP、Wald-IC 在保留正确标定的属性方面的表现略优于 Wald-XPD 方法，但在所有模拟条件下，Wald-XPD 方法在修正错误标定的属性方面的表现均优于另外两种方法。(4)实证数据分析的结果表明，Wald-XPD 方法修正后的 **Q** 矩阵与原始数据有最优的拟合度。

在本研究操纵的 5 种因素中，项目质量对 GDI、Hull、Wald-IC、Wald-XPD 方法表现的影响较大，样本量和测验长度也对 4 种修正方法的表现有一定的影响。出现这种现象的原因可能是，项目质量越高、样本量越大以及测验长度越长，被试观察作答反应矩阵中包含的关于 CDM 中未知参数的信息越多，

因此，以上 4 种方法的表现也就越好。与以往研究类似(Kang et al., 2019; Ma & de la Torre, 2020; Nájera et al., 2021)，本研究同样认为属性分布对于 GDI、Hull、Wald-IC、Wald-XPD 方法在 TNR 指标上的表现有细微的影响。出现这种现象的原因可能是，当属性服从均匀分布时所有可能属性掌握模式分布的概率是相等的，即被试观察作答反应矩阵中包含的关于结构参数的信息是一样的。当属性服从高阶分布时，属性之间存在一定的关联性，使某些属性掌握模式分布的概率可能会比较高，另外一些属性掌握模式分布的概率会比较低，故被试观察作答反应矩阵中包含的结构参数的信息量较少。于是，当属性服从均匀分布时，4 种方法在各个指标上的表现略优。**Q** 矩阵错误设定的比例对 GDI、Wald-IC、Hull 方法表现的影响较大，随着 **Q** 矩阵错误设定比例的降低，它们能够获得更高的 QRR、TPR 和 TNR 值，这与已有研究结果一致(Ma & de la Torre, 2020; Nájera et al., 2021)。然而，**Q** 矩阵错误设定的比例对 Wald-XPD 方法表现的影响则相对较小，结合 Wald-XPD 在 TNR 指标上的表现，本研究认为可能是 Wald-XPD 在迭代结束前的循环中能够有效修正 **Q** 矩阵错误标定的属性。

此外，研究结果表明，Wald-XPD 方法在 TPR 和 TNR 指标上与 Wald-IC、HullP 方法的表现不同。在 TPR 指标上，Wald-XPD 受项目质量低的影响明显，在 TNR 指标上，Wald-IC 和 HullP 受项目质量低以及测验长度短这两种因素的影响明显。TPR 指标数值低，说明 **Q** 矩阵修正方法倾向于修改正确标定的属性，TNR 数值低则说明 **Q** 矩阵修正方法修改错误标定属性的能力弱。综合 TPR、TNR 两个指标可知，虽然 Wald-XPD 方法在项目质量较低的条件下能够较为有效地修正错误标定的属性，但是存在过度修改正确标定属性的倾向。换言之，Wald-XPD 方法虽然提高了 **Q** 矩阵修正的表现，但是在项目质量较低的条件下，有可能会错误地修正了正确标定的 q 元素。Wald-IC 以及 HullP 虽然在项目质量较低的条件下不存在过度修改正确标定属性的倾向，但却无法有效修正错误标定的属性，尤其是 HullP 方法。所以，本研究建议使用 **Q** 矩阵修正方法时，需要注意项目质量，若项目质量较低，可以结合多种修正方法、参考专家意见进而获得准确的 **Q** 矩阵。

本研究采用 C++语言编写 XPD 矩阵，在一定程度上能够提高 **Q** 矩阵修正的速度，但是，由于 Wald-XPD 方法考虑模型中的全部参数且采用迭代的方式进行，在一些条件下可能耗时较长。例如，Wald-XPD 方法最短的平均用时是 12.50 s，最长的平均时间需要 746.01 s。Wald-XPD 方法在各个模拟条件下的平均运行时间见表 5。

6.2 研究展望

本研究提出的 Wald-XPD 方法在 **Q** 矩阵修正中有较好的表现，但仍存在一些不足之处，值得后续研究者进一步探讨。(1)虽然 Wald-XPD 统计量有明确的渐近分布(χ^2 分布)，不需要像 GDI 类方法那样人为地确定一个截止值，但限于研究目的和篇幅本文仅在 0.05 显著性水平上对于 Wald-XPD 统计量的表现进行了显著性检验，未来研究者可以进一步探讨不同的显著性水平对于 Wald-XPD 统计量表现的影响。(2)本研究仅以完整信息矩阵中的 XPD 矩阵构建 Wald 统计量进行 **Q** 矩阵修正，除了 XPD 矩阵之外，研究者还可以将其他完整信息矩阵构建的 Wald 统计量用于 **Q** 矩阵修正，如 Liu 等人(2021)提出改进的观察信息矩阵以及三明治信息矩阵。不同类型的完整信息矩阵构建的 Wald 统计量在 **Q** 矩阵修正中的表现也值得进一步研究。(3)本研究仅在 G-DINA 模型下对 **Q** 矩阵修正方法进行了对比研究，G-DINA 模型适用于 0-1 计分的测验情景，但在心理与教育测验中存在较多的多级计分数据。研究者们开发了很多能用于多级计分的 CDM，如多级计分 GDM (von Davier, 2008)，研究者可以将 Wald-XPD 方法拓展到多级计分模型中，并考察其在多级计分模型中的表现。(4)本研究在考察新提出的 Wald-XPD 方法的表现时，仅与一次修正的 GDI、Wald-IC 方法进行了比较，研究者也认为 GDI、Wald-IC 方法可以迭代进行，如迭代 GDI 方法(Nájera et al., 2020)。此外，还有其他迭代修正的方法，如迭代修正序列搜索(Terzi & de la Torre, 2018)等，研究者也可以尝试将这些方法与 Wald-XPD 方法进行比较。(5) Wang 等人(2020)评估了在 **Q** 矩阵部分已知的情况下，GDI 和 Wald-IC 方法在估计新项目的 **q** 向量中的表现。基于此，未来研究者可以在 **Q** 矩阵部分已知的情况下进一步评估 Wald-XPD 方法估计 **Q** 矩阵的表现，并与已有的 **Q** 矩阵估计方法，如 ICC-IR 方法(汪大勋，高旭亮，蔡艳 等，2018)、似然比 D^2 方法(喻晓锋 等，2015)、非参数 **Q** 矩阵校准(Lim & Drasgow, 2017)、两阶段搜索算法(Feng, 2013)、似然比检验(Wang et al., 2020)等方法进行比较。

表 5　Wald-XPD 方法在各模拟条件下的平均运行时间(s)

模拟条件	*AD*	*QM*	*IQ*	*N*	*JK*	时间
1	均匀分布	0.15	0.4	500	4	476.16
2	高阶分布	0.15	0.4	500	4	195.68
3	均匀分布	0.15	0.4	500	8	706.40
4	高阶分布	0.15	0.4	500	8	654.93
5	均匀分布	0.15	0.4	1000	4	302.90
6	高阶分布	0.15	0.4	1000	4	746.01*
7	均匀分布	0.15	0.4	1000	8	505.79
8	高阶分布	0.15	0.4	1000	8	320.67
9	均匀分布	0.15	0.6	500	4	68.17
10	高阶分布	0.15	0.6	500	4	67.66
11	均匀分布	0.15	0.6	500	8	54.11
12	高阶分布	0.15	0.6	500	8	81.36
13	均匀分布	0.15	0.6	1000	4	21.35
14	高阶分布	0.15	0.6	1000	4	90.22
15	均匀分布	0.15	0.6	1000	8	56.11
16	高阶分布	0.15	0.6	1000	8	113.40
17	均匀分布	0.15	0.8	500	4	12.93
18	高阶分布	0.15	0.8	500	4	21.20
19	均匀分布	0.15	0.8	500	8	23.63
20	高阶分布	0.15	0.8	500	8	46.23
21	均匀分布	0.15	0.8	1000	4	12.97
22	高阶分布	0.15	0.8	1000	4	12.50#
23	均匀分布	0.15	0.8	1000	8	48.36
24	高阶分布	0.15	0.8	1000	8	32.42
25	均匀分布	0.3	0.4	500	4	114.85
26	高阶分布	0.3	0.4	500	4	223.68
27	均匀分布	0.3	0.4	500	8	750.26
28	高阶分布	0.3	0.4	500	8	310.86
29	均匀分布	0.3	0.4	1000	4	163.41
30	高阶分布	0.3	0.4	1000	4	226.47
31	均匀分布	0.3	0.4	1000	8	510.00
32	高阶分布	0.3	0.4	1000	8	696.73
33	均匀分布	0.3	0.6	500	4	63.20
34	高阶分布	0.3	0.6	500	4	111.59
35	均匀分布	0.3	0.6	500	8	64.36
36	高阶分布	0.3	0.6	500	8	111.91
37	均匀分布	0.3	0.6	1000	4	61.61
38	高阶分布	0.3	0.6	1000	4	77.84
39	均匀分布	0.3	0.6	1000	8	95.48
40	高阶分布	0.3	0.6	1000	8	152.57
41	均匀分布	0.3	0.8	500	4	45.05
42	高阶分布	0.3	0.8	500	4	12.75
43	均匀分布	0.3	0.8	500	8	22.22
44	高阶分布	0.3	0.8	500	8	72.12
45	均匀分布	0.3	0.8	1000	4	15.34
46	高阶分布	0.3	0.8	1000	4	25.56
47	均匀分布	0.3	0.8	1000	8	54.40
48	高阶分布	0.3	0.8	1000	8	190.39

注：*为 Wald-XPD 方法在模拟条件下的最长运行时间, #为 Wald-XPD 方法在模拟条件下的最短运行时间。

参 考 文 献

Chen, F., Liu, Y., Xin, T., & Cui, Y. (2018). Applying the M_2 statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies. *Frontiers in Psychology, 9*, Article 1875.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*(4), 277–293.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

Feng, Y. (2013). *Estimation and Q-matrix validation for diagnostic classification models* (Unpublished doctoral dissertation). University of South Carolina, Los Angeles, America.

Gu, Y., Liu, J., Xu, G., & Ying, Z. (2018). Hypothesis testing of the Q-matrix. *Psychometrika, 83*(3), 515–537.

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 42*, 49–56.

Kang, C. H., Yang, Y. K., & Zeng, P. H. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement, 43*(7), 527–542.

Li, J., Mao, X., & Wei, J. (2022). A simple and effective new method of Q-matrix validation. *Acta Psychologica Sinica, 54*(8), 996–1008.

[李佳, 毛秀珍, 韦嘉. (2022). 一种简单有效的 Q 矩阵修正新方法. *心理学报, 54*(8), 996–1008.]

Li, J., Mao, X., & Zhang, X. (2021). Q-matrix estimation (validation) methods for cognitive diagnosis. *Advances in Psychological Science, 29*(12), 2272–2280.

[李佳, 毛秀珍, 张雪琴. (2021). 认知诊断 Q 矩阵估计(修正)方法. *心理科学进展, 29*(12), 2272–2280.]

Li, X., & Wang, W. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28–54.

Lim, Y., & Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix in cognitive diagnosis. *Multivariate Behavioral Research, 52*(5), 562–575.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548–564.

Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement*, *43*(5), 402–414.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*(1), 3–26.

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology, 72*(1), 18–37.

Liu, Y., Xin, T., & Jiang, Y. (2021). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2021.1919048

Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016). An improved method for differential item functioning detection in cognitive diagnosis models: An application of Wald statistic based on observed information matrix. *Acta Psychologica Sinica, 48*(5), 588–598.

[刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信息矩阵的 Wald 统计量. *心理学报, 48*(5), 588–598.]

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253–275.

Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in economics* (pp. 105–142). New York, NY: Academic Press.

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement, 79*(4), 727–753.

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2020). Improving robustness in Q-Matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement, 44*(6), 431–446.

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2021). Balancing fit and parsimony to improve Q-matrix validation. B*ritish Journal of Mathematical and Statistical Psychology, 74*(Suppl 1), 110–130.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics, 43*(1), 88–115.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78–96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. Guilford.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 1–17.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*(3), 506–532.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Terzi, R. (2017). *New Q-matrix validation Procedures* (Unpublished doctoral dissertation). The State University of New Jersey, New Brunswick, America.

Terzi, R., & de la Torre, J. (2018). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education, 5*(2), 248–262.

Tu, D., Cai, Y., & Dai, H. (2012). A new method of Q-Matrix validation based on DINA model. *Acta Psychologica Sinica, 44*(4), 558–568.

[涂冬波, 蔡艳, 戴海琦. (2012). 基于 DINA 模型的 Q 矩阵修正方法. *心理学报, 44*(4), 558–568.]

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

Wang, D., Cai, Y., & Tu, D. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix. *Multivariate Behavioral Research.* Advance online publication. https://doi.org/10.1080/00273171.2020.1746901

Wang, D., Gao, X., Cai, Y., & Tu, D. (2018). A new Q-matrix estimation method: ICC based on ideal response. *Journal of Psychological Science, 41*(2), 466–474.

[汪大勋, 高旭亮, 蔡艳, 涂冬波. (2018). 一种非参数化的 Q 矩阵估计方法: ICC-IR 方法开发. *心理科学, 41*(2), 466–474.]

Wang, D., Gao, X., Cai, Y., & Tu, D. (2020). A method of Q-matrix validation for polytomous response cognitive diagnosis model based on relative fit statistics. *Acta Psychologica Sinica, 52*(1), 93–106.

[汪大勋, 高旭亮, 蔡艳, 涂冬波. (2020). 基于类别水平的多级计分认知诊断 Q 矩阵修正：相对拟合统计量视角. *心理学报, 52*(1), 93–106.]

Wang, D., Gao, X., Han, Y., & Tu, D. (2018). A simple and effective Q-matrix estimation method: From non-parametric perspective. *Journal of Psychological Science, 41*(1), 180–188.

[汪大勋, 高旭亮, 韩雨婷, 涂冬波. (2018). 一种简单有效的 Q 矩阵估计方法开发：基于非参数化方法视角. *心理科学, 41*(1), 180–188.]

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement, 42*(6), 446–459.

Yu, X. F., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology, 73*(Suppl 1), 145–179.

Yu, X., Luo, Z., Qin, C., Gao, C., & Li, J. (2015). Joint estimation of model parameters and Q-matrix based on response data. *Acta Psychologica Sinica, 47*(2), 273–282.

[喻晓锋, 罗照盛, 秦春影, 高椿雷, 李喻骏. (2015). 基于作答数据的模型参数和 Q 矩阵联合估计. *心理学报, 47*(2), 273–282.]

An empirical Q-matrix validation method using complete information matrix in cognitive diagnostic models

LIU Yanlou[1], WU Qiongqiong[2]

([1] *Academy of Big Data for Education; Qufu Normal University, Jining* 273165, *China*)

([2] *School of Psychology, Qufu Normal University, Jining* 273165, *China*)

Abstract

A **Q**-matrix, which defines the relations between latent attributes and items, is a central building block of the cognitive diagnostic models (CDMs). In practice, a **Q**-matrix is usually specified subjectively by domain experts, which might contain some misspecifications. The misspecified **Q**-matrix could cause several serious problems, such as inaccurate model parameters and erroneous attribute profile classifications. Several **Q**-matrix validation methods have been developed in the literature, such as the G-DINA discrimination index (GDI), Wald test based on an incomplete information matrix (Wald-IC), and Hull methods. Although these methods have shown promising results on **Q**-matrix recovery rate (QRR) and true positive rate (TPR), a common drawback of these methods is that they obtain poor results on true negative rate (TNR). It is important to note that the worse performance of the Wald-IC method on TNR might be caused by the incorrect computation of the information matrix.

A new **Q**-matrix validation method is proposed in this paper that constructs a Wald test with a complete empirical cross-product information matrix (XPD). A simulation study was conducted to evaluate the performance of the Wald-XPD method and compare it with GDI, Wald-IC, and Hull methods. Five factors that may influence the performance of **Q**-matrix validation were manipulated. Attribute patterns were generated following either a uniform distribution or a higher-order distribution. The misspecification rate was set to two levels: QM = 0.15 and QM = 0.3. Two sample sizes were manipulated: 500 and 1000. The three levels of IQ were defined as high IQ, $P_j(0) \sim U\,(0, 0.2)$ and $P_j(1) \sim U\,(0.8, 1)$; medium IQ, $P_j(0) \sim U\,(0.1, 0.3)$ and $P_j(1) \sim U\,(0.7, 0.9)$; and low IQ, $P_j(0) \sim U\,(0.2, 0.4)$ and $P_j(1) \sim U\,(0.6, 0.8)$. The number of attributes was fixed at $K = 4$. Two ratios of the

number of items to attribute were considered in the study: $J = 16[(K = 4)\times(JK = 4)]$ and $J = 32[(K = 4)\times(JK = 8)]$.

The simulation results showed the following.

(1) The Wald-XPD method always provided the best results or was close to the best-performing method across the different factor levels, especially in the terms of the TNR. The HullP and Wald-IC methods produced larger values of QRR and TPR but smaller values of TNR. A similar pattern was observed between HullP and HullR, with HullP being better than HullR. Among the **Q**-matrix validation methods considered in this study, the GDI method was the worst performer.

(2) The results from the comparison of the HullP, Wald-IC, and Wald-XPD methods suggested that the Wald-XPD method is more preferred for **Q**-matrix validation. Even though the HullP and Wald-IC methods could provide higher TPR values when the conditions were particularly unfavorable (e.g., low item quality, short test length, and low sample size), they obtain very low TNR values. The practical application of the Wald-XPD method was illustrated using real data.

In conclusion, the Wald-XPD method has excellent power to detect and correct misspecified q-entry. In addition, it is a generic method that can serve as an important complement to domain experts' judgement, which could reduce their workload.

Keywords cognitive diagnostic models, **Q**-matrix, XPD information matrix, Wald test